DOCUMENT RESUME

ED 307 323 TM 013 472

AUTHOR Mayo, Samuel T.
TITLE Helping Teachers Develop Valid and Reliable Assessments: Our Experience in Illinois.
PUB DATE 89
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Content Validity; Criterion Referenced Tests; Elementary School Teachers; Elementary Secondary Education; *Guides; *Inservice Teacher Education; Interrater Reliability; Mastery Tests; School Districts; Scoring; Secondary School Teachers; *Standardized Tests; *Teacher Made Tests; *Test Construction; Test Manuals; Test Reliability; Videotape Recordings
IDENTIFIERS Illinois; Internal Consistency; Paper and Pencil Tests; Split Half Test Reliability

ABSTRACT

While Illinois educational officials developed a uniform assessment manual to accompany standardized tests, a companion handbook was developed by a consultant for use with teacher- and district-made tests. Criterion-referenced or mastery tests were the subject of the handbook. The handbook was designed to assess content validity as well as provide "split-half" and internal consistency information to determine test reliability. Statistical methods incorporated into the handbook utilized classical as well as the simpler, criterion-referenced methods. A training videotape is designed to accompany the handbook, and it includes short student performances of five skills; the tape is designed for use during inservice training as a means of illustrating actual performance upon which the trainees can hone their rating abilities. Specific topics covered in the handbook include establishment of testing objectives, development of paper-and-pencil tests, performance or product assessments, establishment of levels of reliability of paper-and-pencil tests and performance ratings, objectives and assignments related to performance assessments, rating performance, rating scales and checklists and scoring, and interrater reliability. A copy of the handbook's detailed table of contents is appended. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307323

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SAMUEL T. MAYO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

HELPING TEACHERS DEVELOP VALID AND RELIABLE ASSESSMENTS: OUR EXPERIENCE IN ILLINOIS

Samuel T. Mayo
Professor Emeritus, Loyola University of Chicago

and

Consultant, Glenview, Illinois

TM013472

BEST COPY AVAILABLE

Helping Teachers Develop Valid and Reliable Assessments: Our Experience in Illinois

Samuel T. Mayo

Professor Emeritus, Loyola University of Chicago

and

Consultant, Glenview, Illinois

BACKGROUND

The purpose of the project was to make validity and reliability methods accessible to all Illinois teachers for their local assessments. We hoped to provide an intermediate ground between the standard textbooks in the field and everyday classroom practice in measurement. We are involved because we wanted to guarantee that any school, including our own, could continue to use teacher-made tests and local performance assessment as components of its Learning Assessment Plan as response to the Illinois reform package. We also wanted to develop a manual which could be used by any classroom teacher of any subject and any grade (as well as by any administrator) to help meet Illinois' high standards for locally-designed, criterion-referenced, achievement tests of cognitive and psychomotor skills.

All of the 50 states are engaged in school reform and are developing assessment procedures to measure progress toward reform goals. The Illinois Legislature has mandated that each school district submit not only periodic assessment data but in addition evidence of acceptable validity and reliability of local assessment methods. While the State was in the process of publishing a uniform Assessment Handbook to provide a general guide to districts, we conceived the need for a companion handbook, or manual. Such a manual would provide training material in easy techniques for calculating estimates of validity and reliability of both cognitive tests and rating scales of performance.

Our final training package consisted of a Manual and a Videotape. These were produced by a team of five persons. The Project Director and Assistant Project Director were Leslie R. Wilson and Barbara C. Scherbarth, respectively, from the faculty of Homewood-Flossmoor Community High School in South Cook County, Illinois. One Consultant, Samuel T. Mayo, was from the Chicago Area, while Henry M. Brickell and Regina H. Paul were from Policy Studies in Education in Great Neck, New York. The whole team was convened in Chicago on two occasions in order to brainstorm. Otherwise, we handled much by mail and by telephone. Later, the three of us from Chicago Metro met to finish the writing and printing of the Manual. The videotape was produced in a short period in Homewood-Flossmoor Community High School, using its students as performers. During writing, editing, and revisions, several dozen school faculty and administrators acted as readers of the manuscripts.

One of the early tough decisions the team had to make was to select those methods from the literature of estimating validity and reliability which (1)match the practical situations in which teachers found themselves; and (2)can be adapted into manageable computing routines.

Selection of validity methods was much easier than selection of reliability methods. Of the four types of validity estimation in the literature: (1)Content, (2)Construct, (3)Concurrent, and (4)Predictive; we ruled out in short order the latter three as inappropriate for local assessment situations. Therefore, we went with content validity exclusively in our manual. Exhibit B shows the checklist which we recommend. It gathers data upon two aspects: (1)Content Match; and (2)Intellectual Process Match (for objective test items).

Of the four distinct kinds of test reliability: Test-Retest, Parallel Forms, Split-Half, and Internal Consistency, we ruled out early on the test-retest and parallel forms as inappropriate for teacher-made and district-made tests. We considered each of these methods too intrusive to an instructional program and unwieldy in an individual K-12 classroom. These methods are appropriate for standardized tests. We did, however, rule them out reluctantly. The two methods we kept for the manual, split-half and internal consistency, may be thought of as compromises, because they are practical and because they can be considered as estimates of the first two. This is clear in the split-half case. The half-tests operate as parallel forms of the test given on the same day. Under any circumstances, reliability indexes for paper-and-pencil tests at least give (albeit roughly) some indication of whether we are on target in making accurate assessments.

When faced with the practical problem of finding and using appropriate statistical methods to monitor reliability of criterion-referenced (or mastery) tests, we found ourselves on two horns of a dilemma: (1)on the one hand, if we had used any of the four reliability methods previously mentioned (the "classical" methods for norm-referenced tests), we would have run the risk of obtaining misleading information from the data, due to an assumption about relatively wide variability needed in test scores; and (2)reliability measures for criterion-referenced testing are not well-developed, practically speaking. The more elaborate measures are difficult to calculate and are controversial; the simpler ones are, perhaps, too simple and sometimes difficult to interpret. We ended up by compromising between the two horns and using two of the classical methods and some of the simpler, criterion-referenced methods.

ORGANIZATION OF THE MANUAL

Exhibit A shows the Table of Contents of the Manual. The first chapter, MAKING SURE, includes general introductory sections on objectives, paper-and-pencil test assessment and performance assessment.

The second chapter, ENSURING RELIABILITY, includes the more technical sections of the manual. Reliability for cognitive tests and reliability for performance ratings are treated separately. The specific reliability indexes we used were:

Paper and Pencil Testing

(1) Hoyt's method for internal consistency of test items, one occasion
(2) Rulon's formula for split-half reliability of a test, one occasion
(3) Mastery consistency by phi coefficient using counts by split-half, one occasion
(4) Mastery consistency by Kappa coefficient using counts by split-half, one occasion

Performance Ratings

(1) Single-rater, two ratings over time, using Rulon's formula
(2) Two-raters, one occasion, using Rulon's formula
(3) Single-rater, two occasions, by Rater Agreement (%) from Ill. State Bd. of Ed.
(4) Two raters, one occasion, by Rater Agreement (%) from Ill. State Bd. of Ed.
(5) Multiple raters, one occasion, Inter-rater reliability

The sources in the literature for the indexes above are shown in Exhibit C. A worked out example of inter-rater reliability is shown in Exhibit D.

The third chapter, PERFORMANCE ASSESSMENT, provides instruction in developing rating scales, checklists, and scoring procedures for performance assessment. Also, sample rating forms used to record performance ratings from 10 different professional organizations and agencies are included.

THE VIDEOTAPE

Our training tape includes short student performances of five skills:

throwing a 16" softball

shooting a lay-up

dancing: one combination

acting: characterizing old age

speaking: delivering an original speech

The tape is intended to be used during inservice training as a means of illustrating actual performances upon which the trainees can hone their own rating abilities.

RESULTS FROM FIELD TESTING OF THE TRAINING PACKAGE

We have made several kinds of field presentations since publication of our package in the Summer of 1988. We ran two workshops for public school teachers and administrators in South Cook County at the Educational Service Center there. We presented another workshop for both public and private school art faculty and staff from North and West metro Chicago at the School of the Art Institute of Chicago. A presentation was made to educators in downstate Illinois at a meeting of Illinois Association for Research and Evaluation, held in Peoria.

The workshops have shown us that the technical methods for computing reliability can indeed be easily performed by any teacher or administrator with a good feeling and without fear. After completing a workshop, the manual can be easily used on one's own, since we have encouraged persons to make up their own computing work sheets to facilitate entering data and processing the separate computing steps. Computations can be done with paper and pencil and an inexpensive pocket calculator and WITHOUT A COMPUTER. This is not to say that a computer would not be easier, but as of now, the computer programs have yet to be written.

We feel that we have accomplished what we set out to do: provide an accessible route toward establishing the validity and reliability of teacher-made tests and a vehicle for incorporating and monitoring performance assessment in local assessment plans.

Exhibit A

TABLE OF CONTENTS

Page

Foreword i

Introduction vii

CHAPTER ONE: MAKING SURE

Establishing Good Objectives 1

Why Is It So Important to Establish Good Objectives? 1
What Does a Good Objective Look Like? 1
How to Check How Good Your Objectives Are 4

Establishing Good Paper-and-Pencil Testing 5

Establishing Good Questions for Paper-and-Pencil Testing 6

How Many Good Questions 7

Establishing Validity of Paper-and-Pencil Testing 8

What Is Content Validity? 8
How to Establish Content Validity 9

Establishing Reliability of Paper-and-Pencil Testing 10

What Is Reliability? 10
Preparing to Determine Consistency 11
How to Calculate Consistency 12
How Much Consistency Is Enough? 13

Establishing Good Performance or Product Assessments 15

Assessing the Performance or the Product--or Both 15
Recording the Performance Before Assessing It 16
Building Archives of Performances and Products 16
Measuring Performances 17
Considerations about Judging 17

Establishing Good Rating Scales 18

What Will the Scale Ask the Teacher to Judge? 19
How Many Points Will the Scale Have? 19
How Will the Points Be Labelled? 20
How Many Good Rating Scales? 21

Establishing Validity of Performance or Product Assessments 22

How to Establish Content Validity 22

Establishing Reliability of Performance or Product Assessments 23

Training the Teachers to Use Rating Scales 24
Calculating Reliability 24
Reliability of Ratings by One Teacher 24
Reliability of Ratings by Two or More Teachers 26

CHAPTER TWO: ENSURING RELIABILITY

Establishing Reliability for Paper and Pencil Testing 31

Theoretical Considerations 31

Consistency Methods 32

Internal Consistency Reliability 32

Split-Half Reliability for a Test ("Long" Method) 34

Reliability Index 36

Split-Halves: Mastery Consistency Methods 38

Consistency Indexes for Criterion-Referenced (Mastery) Tests 40

Establishing Reliability of Performance Ratings 43

Performance Conditions 43

Selecting Raters for Performance Ratings 43

Training Raters for Performance Ratings 43

Common Errors in Rating 44
Single-Rater Reliability 46
Two-Rater Reliability 47
Multi-Rater Reliability 49
Worksheet 53
Chapter Bibliography 54

CHAPTER THREE: PERFORMANCE ASSESSMENT

Performance Assessment: Objectives and Assignments 59
Characteristics of Good Assignments 59
Process for Designing Assignments 60
Examples 60
Rating Performance 66
Choosing the Right Criteria for Judging Performance 66
Example Criteria 67
Rating Scales, Checklists, and Scoring 70
Holistic vs. Analytic Scoring 70
Directory for Sample Rating Forms 72
Sample Rating Forms 75
Art 75
Dance 76
Drama 80
Music 85
Speech 97
Composition 102

Developing Inter-Rater Reliability: Issues to Consider 103
Standard Score 103
Standard Deduction 104
Recording Performance Ratings 105
Student Profiles 105
Objective Profiles 106
Class Roster Profiles 106
Chapter Bibliography 108

PROJECT NOTES 110
ANNOTATED BIBLIOGRAPHY 115
Special Thanks 119

How To Establish Content Validity

Here is the most straightforward way to establish the content validity of the four-question test for each objective once you have written the questions:

1. Find two other teachers who are familiar with the content of the objective. (If they have taught the objective themselves, so much the better.) Ask them to serve as judges.

2. Ask each judge to examine each question independently to see whether it matches:

   -- *the exact content of the objective.*
   -- *the intellectual process called for by the objective.*

   Ask each judge to record ratings on a sheet like this:

Objective:_____		
Question Number	Content Match	Intellectual Process Match
1	Yes No	Yes No
2	Yes No	Yes No
3	Yes No	Yes No
4	Yes No	Yes No

3. Tally the judges' ratings. Any "no" vote makes that question not valid.

4. If any questions were judged not valid, write some more questions and have those new questions judged. Keep writing and having questions judged until you have four questions that receive all "yes" votes by both judges.

5. Ask each judge to look independently at the four questions as a group and determine whether the four questions taken together represent the full range of learning called for by the objective. Both judges must agree that they do.

6. If the judges do not agree that the learning is well represented, write some new questions on whatever gaps the judges can help you identify, have those questions judged on content and intellectual process match, and have the improved group of four questions judged for representativeness. Keep writing and judging until a group of four questions is approved as representative by both judges.

You now have a four-question test that has content validity for an objective. This is a high standard to set: two independent judges in complete agreement. Validity of your test is that important if students are to be judged fairly.

Exhibit C

References for Formulas

Cognitive Tests	Performance Ratings	Source
	Single Rater, Two Occasions; or Two Raters, One Occasion (Percent Consistent Ratings)	1. Write on, Illinois!: A User's Guide to Scoring Student Essays. Volume II, Illinois State Board of Education, 1987.
	Multiple Raters, One Occasion (Inter-Rater Reliability)	2. Ebel, Robert L. and David A. Frisbie. Essentials of Educational Measurement (3rd. ed.). Englewood Cliffs, NJ: Prentice-Hall, 1979, pp. 232-284.
Internal Consistency of Test Items (One Occasion)		3. Hoyt, Cyril J. "Test Reliability Estimated by Analysis of Variance." Psychometrika, 1941, 3:153-160.
Split-Half Reliability of a Test (One Occasion)	Single Rater, Two Occasions; or Two Raters, One Occasion	4. Rulon, P.J. "A Simplified Procedure for Determining the Reliability of a Test by Split-Halves." Harvard Educational Review, 1939, 9:99-103.
Mastery Consistency of Split-Halves of a Test, One Occasion (Phi Coefficient)		5. Swezey, Robert W. Individual Performance Assessment: An Approach to Criterion-Referenced Test Development. Reston, VA: Reston Publishing Co., 1981, pp. 145-147.
Mastery Consistency of Split-Halves of a Test, One Occasion (Cohen's Kappa)		6. Nitko, Anthony J. Educational Tests and Measurement: An Introduction. Orlando, FL: Harcourt Brace Jovanovich, 1983, pp. 406-407.

Exhibit D

Computing Diagram for Reliability of Multiple-Raters of Multiple-Students

Rating Scores and Totals for Original Data

Student No.	RATERS A	B	C	D	E	Student Totals
1	3	3	3	3	4	16
2	3	3	3	3	5	17
3	3	3	4	3	4	17
4	4	3	4	3	5	19
5	3	3	4	3	4	17
6	1	1	2	2	3	9
7	3	3	3	3	4	16
8	5	4	5	4	5	23
9	4	3	5	4	5	21
10	4	3	4	4	5	20
11	4	3	5	4	5	21
12	4	3	4	4	5	20
13	2	1	3	3	3	12
14	2	1	3	3	3	12
15	1	1	2	2	3	9
16	2	1	2	2	3	10
17	1	1	2	1	2	7
18	3	3	4	3	4	17
19	3	3	4	4	4	18
20	2	2	3	2	3	12
Rater Totals	57	48	69	60	79	313

Terms of Reliability Equation

n = 20, Number of Students

K = 5, Number of Raters

$$A = \boxed{\frac{\text{Sum of Squares of Student Totals}}{n}} - \boxed{\frac{\text{Square Sum of Student Totals}}{n^2}} = \boxed{\frac{5,307}{20}} - \boxed{\frac{97,969.00}{400}} = 265.35 - 244.9225 =$$

A = 20.4275

$$B = \boxed{\frac{\text{Sum of Squares of Student Ratings}}{n}} - \boxed{\frac{\text{Sum of Squares of Rater Totals}}{n^2}} = \boxed{\frac{1,101}{20}} - \boxed{\frac{20,155}{400}} = 55.05 - 50.3875 =$$

B = 4.6625

The term k/(k-1) in our case is 5/4=1.25. In the brackets, B/A is 4.6225/20.4275=.2282. Next, 1-.2282=.7718. Finally, 1.25 multiplied by .77.8=.96, the Reliability for our 100 ratings of 20 students by 5 raters. This kind of reliability is frequently referred to as "Inter-Rater Reliability."

Reliability Calculations

$$\text{Reliability} = \frac{k}{k-1}\boxed{1 - \frac{B}{A}} = \frac{5}{4}\boxed{1 - \frac{4.6625}{20.4275}} = 1.25\,(1 - .2282) = 1.25\,(.7718)$$

Reliability = .96